ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΠΟΙΟΤΗΤΑΣ ΑΕΡΑ ΣΕ ΜΕΓΑΛΕΣ ΠΟΛΕΙΣ ΑΞΙΟΠΟΙΩΝΤΑΣ ΔΕΔΟΜΕΝΑ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Μεταπτυχιακή Διπλωματική Εργασία

Πολυχρόνης Χαριτίδης

ΕΠΙΒΛΕΠΩΝ/ΟΝΤΕΣ:

Παναγιώτα Τσομπανοπούλου

Ελευθέριος Τσουκαλάς

Μιχαήλ Βασιλακόπουλος

Βόλος 2018

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΠΟΙΟΤΗΤΑΣ ΑΕΡΑ ΣΕ ΜΕΓΑΛΕΣ ΠΟΛΕΙΣ ΑΞΙΟΠΟΙΩΝΤΑΣ ΔΕΔΟΜΕΝΑ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Μεταπτυχιακή Διπλωματική Εργασία

Πολυχρόνης Χαριτίδης

ΕΠΙΒΛΕΠΩΝ/ΟΝΤΕΣ:

Παναγιώτα Τσομπανοπούλου

Ελευθέριος Τσουκαλάς

Μιχαήλ Βασιλακόπουλος

Βόλος 2018

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AIR QUALITY PREDICTION IN LARGE CITIES UTILIZING SOCIAL MEDIA DATA

Master Thesis

Polychronis Charitidis

Supervisor/s:

Panagiota Tsompanopoulou

Eleftherios Tsoukalas

Michael Vasilakopoulos

Volos 2018

ACKNOWLEDGEMENTS

I would like to thank my advisors Eleftherios Tsoukalas, Michael Vasilakopoulos and especially Panagiota Tsompanopoulou for their help and guidance throughout this work.

In addition, I would like to thank my colleagues Lefteris Spyromitros and Akis Papadopoulos from CERTH-ITI for their valuable help. It is due to their inspiration and continuous encouragement that I was able to successfully complete this work.

Also I would like to thank my family for their continuous support throughout these years.

Part of this work was conducted for hackAIR[1] research project and is partially funded by the European Commission under the contract number H2020-688363

[1] http://www.hackair.eu/

ΠΕΡΙΛΗΨΗ

Η ατμοσφαιρική ρύπανση είναι η τέταρτη αιτία θανάτου παγκοσμίως. Ενώ οι θάνατοι που σχετίζονται με την ατμοσφαιρική ρύπανση αφορούν κυρίως παιδιά και ηλικιωμένους, οι θάνατοι αυτοί συντελούν στην οικονομική υποβάθμιση του εργατικού δυναμικού. Η απώλεια της ζωής είναι τραγικό γεγονός. Το κόστος στην οικονομία σημαντικό. Μια από τις σημαντικότερες δράσεις καταπολέμησης των αρνητικών επιπτώσεων της ατμοσφαιρικής ρύπανσης είναι η ενημέρωση και ευαισθητοποίηση των ανθρώπων σχετικά με την ρύπανση του αέρα και η παροχή συνεχής ροής πληροφοριών ποιότητας αέρα. Ωστόσο, το υψηλό κόστος εγκατάστασης και συντήρησης των σταθμών μετρήσεων οδηγούν στην δημιουργία αραιών δικτύων παρακολούθησης τα οποία παρέχουν μετρήσεις σε συγκεκριμένα σημεία μεγάλων πόλεων. Αυτό έχει ως αποτέλεσμα οι πολίτες μικρότερων αστικών πόλεων και υποανάπτυκτων περιοχών να μην έχουν πρόσβαση σε πληροφορίες σχετικές με την ποιότητα του αέρα. Συνεπώς, υπάρχει η ανάγκη για μια αποτελεσματική αλλά οικονομική λύση που αξιοποιεί την καθημερινή τεχνολογία.

Αυτή η μεταπτυχιακή εργασία ερευνά το κατά πόσο είναι εφικτή η εκτίμηση των τρεχουσών συνθηκών ποιότητας αέρα σε πόλεις χωρίς επίσημη πληροφορία από σταθμούς μέτρησης, με βάση την στατιστική ανάλυση της δραστηριότητας στα κοινωνικά δίκτυα. Για τον σκοπό αυτό, σχεδιάσαμε και υλοποιήσαμε ένα σύστημα για την συλλογή και γεωγραφική σήμανση των δημοσιεύσεων στα κοινωνικά δίκτυα και την εφαρμογή μοντέλων μηχανικής μάθησης και μεταφοράς γνώσης. Το σύστημα αυτό παράγει εκτιμήσεις για τα επίπεδα ρύπανσης του αέρα σε πόλεις που δεν έχουν δεδομένα μετρήσεων αξιοποιώντας δεδομένα από κοντινές πόλεις που διαθέτουν πραγματικές μετρήσεις.

Τα πειράματα που έγιναν αγγλικές και αμερικάνικες πόλεις εστιάζοντας στο Twitter, υποδεικνύουν ότι ενώ οι εκτιμήσεις με βάση το Twitter έχουν πολύ υψηλή ακρίβεια, αποδίδουν χειρότερα κατά μέσο όρο σε σχέση με μια απλή χωρική παρεμβολή. Ωστόσο, παρατηρήσαμε ότι ένα μοντέλο που συνδυάζει τις μετρήσεις από την χωρική παρεμβολή και τις μετρήσεις με βάση το Twitter αυξάνει την ακρίβεια στις πιο απομακρυσμένες πόλεις, επισημαίνοντας την αξία του Twitter για την εκτίμηση της ατμοσφαιρικής ρύπανσης και την περεταίρω έρευνα πάνω στο θέμα.

ABSTRACT

Air pollution has emerged as the fourth-leading risk factor for deaths worldwide. While pollution-related deaths mainly strike young children and the elderly, these deaths also result in lost labour income for working-age men and women. The loss of life is tragic. The cost to the economy is substantial. One of the main areas for action to battle the adverse effects of air pollution is raising people's awareness with respect to air quality and providing them access to real-time air quality information. However, the high costs of installation, maintenance and calibration of reference stations result in sparse monitoring networks that provide measurements only in few locations of large cities. As a result, citizens of smaller urban areas and underdeveloped regions, lack accessibility to air quality information. That being said, there is a need of an effective, yet cost-efficient solution that utilizes everyday technology.

This MSc thesis investigates the feasibility of estimating current air quality conditions in cities without official air quality monitoring stations based on a statistical analysis of social media activity. For this purpose, a framework for collecting and geotagging air quality-related social media posts is developed and transfer learning is applied to enable estimations for unmonitored cities using data from monitored nearby cities.

Experiments carried out on English and American cities focusing on Twitter, suggest that while Twitter-based estimates exhibit very high accuracy, they are outperformed on average by simple spatial interpolation. However, we find that a meta-model that combines estimates from spatial interpolation with Twitter-based ones increases accuracy in distantly located cities, highlighting the merits of Twitter-based air quality estimation and motivating further work on the topic.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Air pollution is a major environmental issue that occurs as a result of human activity (e.g. carbon emissions from cars) or natural processes (e.g. volcanic eruptions) and its implications range from damages to food crops and buildings to diseases in animals and humans. According to the World Health Organization (WHO), air pollution is responsible for an estimated 11.6% of all global deaths in 2012, while 92% of world's population live in places where air pollution exceeds WHO guideline limits. Most of these deaths occur due to fine particulate matter (i.e., $PM_{2.5}$, denoting particles with aerodynamic diameter less than 2.5 microns), for which concentrations are especially high in cities of China. Although the problem is more prominent in countries such as China, a large proportion of European populations and ecosystems are also exposed to air pollution levels that exceed the WHO air quality guidelines according to a 2017 report on air quality in Europe from the European Environmental Agency (EEA) [1].

To determine air quality we need to define the material or substance that cause air pollution. An air pollutant is a material in the air that can have adverse effects on humans and the ecosystem. The substance can be solid particles, liquid droplets, or gases. A pollutant can be of natural origin or man-made. Pollutants are classified as primary or secondary. Primary pollutants are usually produced from a process, such as ash from a volcanic eruption. Other examples include carbon monoxide gas from motor vehicle exhaust, or the sulphur dioxide released from factories. Secondary pollutants are not emitted directly. Rather, they form in the air when primary pollutants react or interact. Ground level ozone is a prominent example of a secondary pollutant. Some pollutants may be both primary and secondary: they are both emitted directly and formed from other primary pollutants.

Substances emitted into the atmosphere by human activity include:

- Carbon dioxide (CO_2) – Because of its role as a greenhouse gas it has been described as "the leading pollutant" and "the worst climate pollution". Carbon dioxide is a natural component of the atmosphere, essential for plant life and given off by the human respiratory system. CO_2 currently forms about 410 parts

per million (ppm) of earth's atmosphere, compared to about 280 ppm in pre-industrial times, and billions of metric tons of CO_2 are emitted annually by burning of fossil fuels. CO_2 increase in earth's atmosphere has been accelerating.

- Sulphur oxides (SOx) – particularly sulphur dioxide, a chemical compound with the formula SO_2. SO_2 is produced by volcanoes and in various industrial processes. Coal and petroleum often contain sulphur compounds, and their combustion generates sulphur dioxide. Further oxidation of SO_2, usually in the presence of a catalyst such as NO_2, forms H_2SO_4, and thus acid rain. This is one of the causes for concern over the environmental impact of the use of these fuels as power sources.
- Nitrogen oxides (NOx) – Nitrogen oxides, particularly nitrogen dioxide, are expelled from high temperature combustion, and are also produced during thunderstorms by electric discharge. They can be seen as a brown haze dome above or a plume downwind of cities. Nitrogen dioxide is a chemical compound with the formula NO_2. It is one of several nitrogen oxides.
- Carbon monoxide (CO) – CO is a colourless, odourless, toxic yet non-irritating gas. It is a product of combustion of fuel such as natural gas, coal or wood. Vehicular exhaust contributes to the majority of carbon monoxide let into our atmosphere. It creates a smog type formation in the air that has been linked to many lung diseases and disruptions to the natural environment and animals.
- Volatile organic compounds (VOC) – VOCs are a well-known outdoor air pollutant. They are categorized as either methane (CH4) or non-methane (NMVOCs). Methane is an extremely efficient greenhouse gas which contributes to enhanced global warming. Other hydrocarbon VOCs are also significant greenhouse gases because of their role in creating ozone and prolonging the life of methane in the atmosphere. The aromatic NMVOCs benzene, toluene and xylene are suspected carcinogens and may lead to leukaemia with prolonged exposure. 1,3-butadiene is another dangerous compound often associated with industrial use.
- Particulates, alternatively referred to as particulate matter (PM), atmospheric particulate matter, or fine particles, are tiny particles of solid or liquid suspended in a gas. In contrast, aerosol refers to combined particles and gas. Some particulates occur naturally, originating from volcanoes, dust storms, forest and

grassland fires, living vegetation, and sea spray. Human activities, such as the burning of fossil fuels in vehicles, power plants and various industrial processes also generate significant amounts of aerosols. Averaged worldwide, anthropogenic aerosols—those made by human activities—currently account for approximately 10 percent of our atmosphere. Increased levels of fine particles in the air are linked to health hazards such as heart disease, altered lung function and lung cancer. Particulates are related to respiratory infections and can be particularly harmful to those already suffering from conditions like asthma.

- Toxic metals, such as lead and mercury, especially their compounds.
- Chlorofluorocarbons (CFCs) – harmful to the ozone layer; emitted from products are currently banned from use. These are gases which are released from air conditioners, refrigerators, aerosol sprays, etc. On release into the air, CFCs rise to the stratosphere. Here they come in contact with other gases and damage the ozone layer. This allows harmful ultraviolet rays to reach the earth's surface. This can lead to skin cancer, eye disease and can even cause damage to plants
- Ammonia (NH_3) – emitted from agricultural processes. Ammonia is a compound with the formula NH_3. It is normally encountered as a gas with a characteristic pungent odour. Ammonia contributes significantly to the nutritional needs of terrestrial organisms by serving as a precursor to foodstuffs and fertilizers. Ammonia, either directly or indirectly, is also a building block for the synthesis of many pharmaceuticals. Although in wide use, ammonia is both caustic and hazardous. In the atmosphere, ammonia reacts with oxides of nitrogen and sulphur to form secondary particles.
- Radioactive pollutants – produced by nuclear explosions, nuclear events, war explosives, and natural processes such as the radioactive decay of radon.

Secondary pollutants include:

- Particulates created from gaseous primary pollutants and compounds in photochemical smog. Smog is a kind of air pollution. Classic smog results from large amounts of coal burning in an area caused by a mixture of smoke and sulphur dioxide. Modern smog does not usually come from coal but from vehicular and industrial emissions that are acted on in the atmosphere by ultraviolet light

from the sun to form secondary pollutants that also combine with the primary emissions to form photochemical smog.

- Ground level ozone (O_3) formed from NOx and VOCs. Ozone (O_3) is a key constituent of the troposphere. It is also an important constituent of certain regions of the stratosphere commonly known as the Ozone layer. Photochemical and chemical reactions involving it drive many of the chemical processes that occur in the atmosphere by day and by night. At abnormally high concentrations brought about by human activities (largely the combustion of fossil fuel), it is a pollutant, and a constituent of smog.
- Peroxyacetyl nitrate ($C_2H_3NO_5$) – similarly formed from NOx and VOCs.

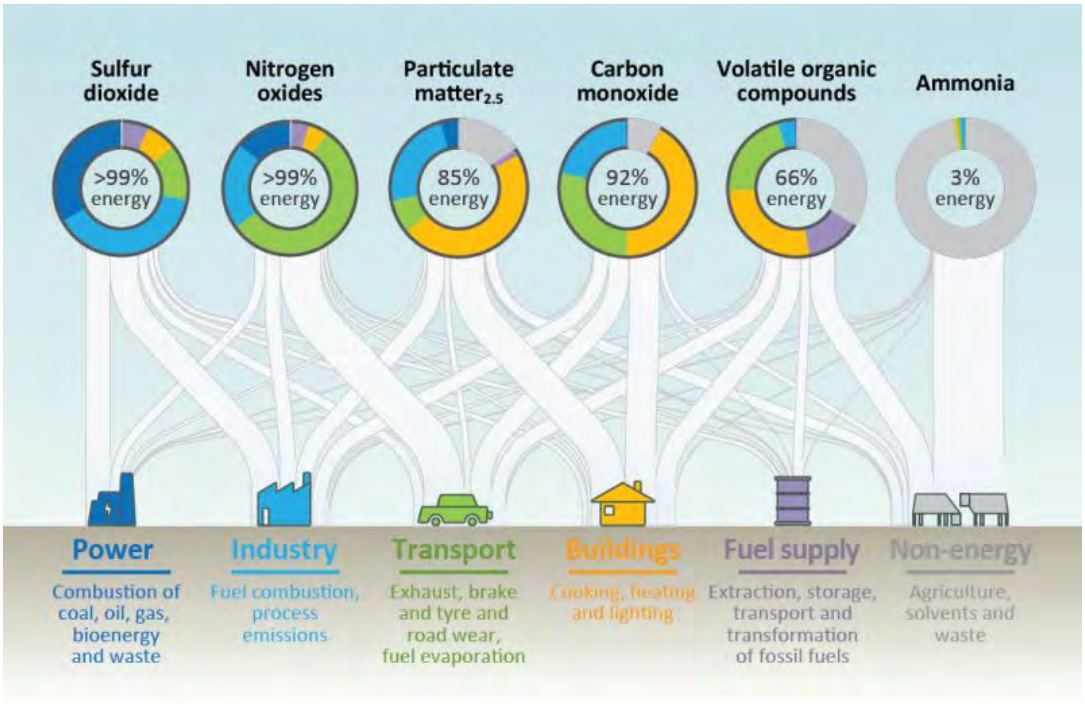

Figure 1: Primary air pollutants[2]

The above air pollutants play crucial role to the status of air quality. Figure 1 shows the primary air pollutants and their sources in 2015.We observe that the transport sector holds a significant share of the global NOx- (over 50%) and is also responsible for about 10% of the energy-related PM-emissions. Concerning the transport section, traffic on roads represents the main emitter of NOx and PM (58% and 73% respectively). Most responsible for these high NOx emissions are heavy and diesel-powered trucks (40-50%).

[2] Source: https://www.sutp.org/en/news-reader/world-energy-outlook-special-report-2016-on-air-pollution-released-9039.html

Nevertheless, the International Energy Agency estimates the health consequences of emissions caused by passenger cars and small load vehicles to be higher as they are mainly used in cities. Concerning Sulphur Oxides, maritime traffic is responsible for most of the emissions.

One of the main areas for action to battle the adverse effects of air pollution is raising people's awareness with respect to air quality and providing them access to real-time air quality information [2], [3]. However, the high costs of installation (€5-30K per monitoring device [4]), maintenance and calibration of reference stations result in sparse monitoring networks that provide measurements only in few locations of large cities. As a result, citizens of smaller urban areas and underdeveloped regions, lack accessibility to air quality information. While low-cost sensors provide a promising alternative means of air quality monitoring in such areas, they are characterized by low robustness and measurement repeatability [4] and, despite low, their cost is not negligible.

In the past, several researchers have tried to address the problems of estimating current air quality in unmonitored locations (spatial prediction) and short-term air quality forecasting (temporal prediction) using statistical approaches that model the relationships between air pollutants and various explanatory variables such as lagged pollutant observations, wind speed, solar radiation, cloud coverage, air temperature, traffic, etc. ([5] contains a detailed review of such methods).

More recently, the rise of online social networks (OSNs) and the wealth of almost real-time information that they provide about a variety of real-world events and phenomena, has motivated the development of air quality prediction methods that are based on a statistical analysis of the publicly available OSN content. That line of work, is based on the view of OSN users acting as "social sensors" [6] and builds upon previous successes on detecting and tracking real-world events (e.g. flu outbreak detection and tracking [7], earthquake detection [8], wildlife roadkill monitoring [9], etc.) based on a statistical analysis of the content posted on these platforms.

In this dissertation we present the first, to the best of our knowledge, attempt to perform OSN-based city-level air quality estimations outside China. In particular, my study focuses on cities in the UK and the US that exhibit important differences compared to Chinese cities:

- High air pollution events are less frequent and pronounced,
- UK and US city population and, consequently, the volume of OSN content they generate is considerably smaller.

Our work is the first to use Twitter as data source for air quality estimation. The reason we focus our work on Twitter is because it offers a simple framework where people can post their current status in real time which is something fundamental for our task. Also, Twitter offers a public Application Programming Interface that makes real-time post collection a fairly easy procedure. Twitter has very similar functionality with Sina Weibo[3], a Chinese microblogging website, which is used in every related research on the topic as discussed in **Error! Reference source not found.** section.

Although this imposes the development of a Twitter-oriented data collection and mining pipeline, it makes the proposed method applicable worldwide. Besides that, previous works consider only 24- hour temporal bins, while in our work we also consider 6-hour and 12-hour ones. While a 24-hour granularity is useful for a post-hoc analysis, finer-grained estimates provide actionable information and it is therefore important to evaluate the accuracy under this setting. Moreover, with small exceptions, all previous related works (which will be discussed in the next section) build and evaluate estimation models on the same city, while current study adopts a more realistic setup where models are evaluated only on cities that have not been used for training.

Our approach collects air quality-related tweets from the Twitter API using a set of air quality-related keywords and then estimates the location to which they refer using a state-of-the-art location estimation method [10]. In the sequel, all tweets falling in a particular spatiotemporal bin are pooled together to form a single textual document that is represented using a Bag-of-Words[4] (BoW) scheme. This representation forms the basis for the developed air-quality estimation models. Compared to simpler types of features such as the number of tweets in each spatiotemporal bin and their polarity (i.e. whether they refer to bad or good air quality) that were used in previous works [11], [12], [13] we found that BoW features lead to better results. Finally, our work is the first that recognizes the multi-task nature of spatial air quality prediction and uses multi-task

[3] https://en.wikipedia.org/wiki/Sina_Weibo
[4] https://en.wikipedia.org/wiki/Bag-of-words_model

learning techniques (data pooling, joint feature selection and sample weighting) to build a robust, city invariant model.

Traditional approaches for spatial air quality prediction include spatial interpolation (e.g. Inverse Distance Weighting (IDW) and variations of Kriging [14]), dispersion models [15] and Land Use Regression (LUR) variants [16]. Among these methods, dispersion and LUR are known to generate robust long-term intra-city predictions (when enough data are available) but spatial interpolation is usually preferred for spatially coarser short-term estimations [5]. Therefore, in our study we compare our city-level Twitter-based estimates with those generated by a spatial interpolation method (IDW). Experiments show that models based only on Twitter information provide fairly accurate estimates but perform worse than spatial interpolation. However, when spatial interpolation estimates are carefully combined with Twitter-based ones, better accuracy can be obtained.

It is also worth pointing out the main limitations of my approach upfront. First, the proposed framework does not forecast future air quality values but rather estimates current air quality utilising population reactions in social media. Second, my model is subject to the availability of social media posts and therefore does not apply to remote regions with extremely low social media user populations. Still, this work provides complementary value to existing air quality monitoring approaches and with further improvements can be an integral part in the overall solution to the air pollution problem.

The rest of the thesis is organized as follows. Chapter 2 presents the related work and some different approaches on the problem. Chapter **Error! Reference source not found.** presents the methodology that is followed to estimate current air quality conditions. In Chapter **Error! Reference source not found.**, experimental evaluation is presented. Finally in Chapter 0 I conclude my Master thesis.

CHAPTER 2

RELATED WORK

Apart from the use of OSN's as "social sensors" for real world events like flu outbreaks, earthquakes, etc. that was mentioned in Chapter **Error! Reference source not found.** there are relevant studies to the current line of work. Specifically, methods of this type have so far been applied for the estimation of city-level air quality in China by analysing content posted in Sina Weibo (a Chinese microblogging website) with encouraging results.

Mei et al. [17] focus their research in 108 Chinese cities and collect on average 1380 Sina Weibo posts on a period of one month for each spatiotemporal bin (city and day). A BoW representation of posts is adapted with 100.000 vocabulary size. They employ regression techniques (Ridge regression [18] and Support vector regression [19]) combined with Markov Random Field [20] method to exploit the spatial correlation between air quality information and social media information. To build the model, data 83 cities are used for training. Data from rest 25 cities are used for the evaluation of the model. They report accurate air quality prediction performance with their approach.

Authors in [11] try to make air quality estimations for Beijing. To accomplish this task they collect approximately 120.000 geo-targeted Sina Weibo posts from Beijing over a period of two years. After pre-processing and filtering retweets (reposted or forwarded posts, originally posted by other user) and automated posts created by bots they use topic modelling techniques to identify positive and negative posts that are relevant with air quality. Having this information they use data from one year to train a Gradient Tree Boosting [21] model and evaluate it with data from the second year. According to their correlation analysis, filtered social media messages are strongly correlated to the air quality index and can be used to monitor the air quality dynamics to some extent.

Wang et al [12] gathered a collection of 93 million Sina Weibo posts in 74 Chinese cities. To identify messages relevant to air quality based on keyword matching and topic modelling. They evaluated the reliability of the data filters by comparing message volume per city to air particle pollution rates obtained from the Chinese government.

Additionally, they performed a qualitative study of the content of pollution-related messages by coding a sample of 170 messages for relevance to air quality, and whether the message included details such as a reactive behaviour or a health concern. In their results they report that the volume of pollution-related messages is highly correlated with particle pollution levels.

A more systematic and focused data collection is achieved in [13] as a collection of 112 million excluding retweets and automated posts was gathered. These posts originated from four Chinese cities over a period of two years. In their approach they utilize bigrams representing 2 consecutive Chinese characters in a sentence resulting in 40 million bigrams in their dataset. For each term, the number of posts containing is counted daily and aggregated separately by city according to the user's registered city, the post count is then divided by the number of all posts in that city on that day , resulting in a fraction of the posts containing that term in a particular city and on a particular day. They calculate correlation coefficients for each term city pair, and the final score of each candidate term of a city is the mean of values in four cities. Using the most correlated terms to construct an Air Discussion Index (ADI) for estimating daily particulate matter values based on the content of Weibo posts by summing each term frequency fraction multiplied by the correlation sign. They conclude that in Beijing, the most polluted Chinese city as measured by U.S. Embassy monitor stations, there is a strong correlation between the ADI and measured particulate matter. In other Chinese cities with lower pollution levels, the correlation between post information and air quality values is weaker. Nonetheless, their results show that social media may be a useful proxy measurement for pollution.

CHAPTER 3

METHODOLOGY

Our work aims at producing accurate estimates of current air quality conditions for cities without air quality monitoring infrastructure based on social media activity and measurements from nearby cities. In **Error! Reference source not found.** we determined the pollutants that affect air quality. We also highlighted the reasons we choose Twitter to be the most appropriate social network for our case. In this step we have to define how we measure air quality and which pollutants we take into consideration. Most research listed in **Error! Reference source not found.** section experiment with particulate matter pollution and especially $PM_{2.5}$.

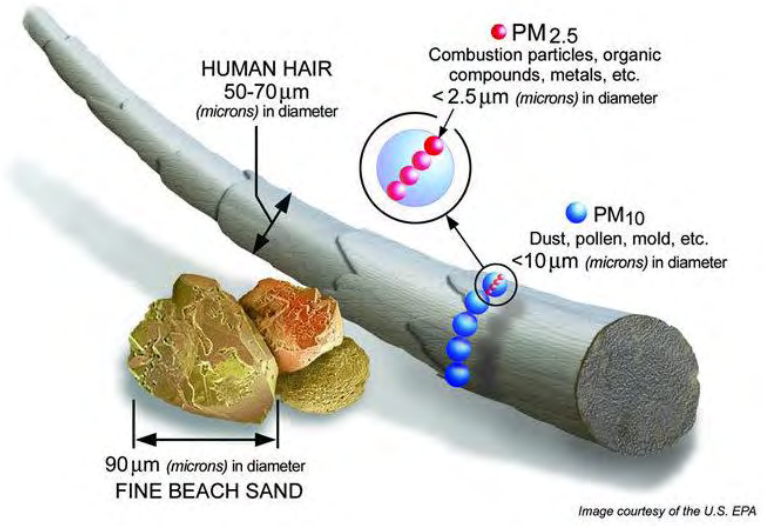

Figure 2: Size of different fine particles[5]

$PM_{2.5}$ refers to atmospheric particulate matter that has a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair. Particles in this category are so small that they can only be detected with an electron microscope. Figure 2 show that they are even smaller than their counterparts PM_{10}, which are particles that are 10 micrometers or less, and are also called fine particles. Since they are so small and light, fine particles tend to stay longer in the air than heavier particles. This increases the chances of humans and animals inhaling them into the bodies. Owing to their minute size,

[5] Source: http://www.irceline.be/en/documentation/faq/what-is-pm10-and-pm2.5

particles smaller than 2.5 micrometers are able to bypass the nose and throat and penetrate deep into the lungs and some may even enter the circulatory system. Studies have found a close link between exposure to fine particles and premature death from heart and lung disease. Fine particles are also known to trigger or worsen chronic disease such as asthma, heart attack, bronchitis and other respiratory problems. A study published in the Journal of the American Medical Association [22] suggests that long-term exposure to $PM_{2.5}$ may lead to plaque deposits in arteries, causing vascular inflammation and a hardening of the arteries which can eventually lead to heart attack and stroke. Scientists in the study estimated that for every 10 micrograms per cubic meter ($\mu g/m^3$) increase in fine particulate air pollution, there is an associated 4%, 6% and 8% increased risk of all-cause, cardiopulmonary and lung cancer mortality, respectively. The American Heart Association has also warned about the impact of $PM_{2.5}$ on heart health and mortality: "Exposure to PM <2.5 μm in diameter ($PM_{2.5}$) over a few hours to weeks can trigger cardiovascular disease-related mortality and nonfatal events; longer-term exposure increases the risk for cardiovascular mortality to an even greater extent than exposures over a few days and reduces life expectancy within more highly exposed segments of the population by several months to a few years."

Having an immediate effect on people's cardiovascular system, $PM_{2.5}$ is a very dangerous pollutant. Consequently, this will make people complain more about their health in conditions where $PM_{2.5}$ is above certain thresholds. Based on that, there will be higher chances of people using Twitter to express their discomfort, symptoms or current air quality status, making $PM_{2.5}$ a suitable and representative pollutant to measure air quality for our needs. **Figure 3** demonstrates some tweets that are relevant to air quality issues. These tweets are perfect examples of people referring on bad air conditions. In this work we focus on estimating $PM_{2.5}$ but it is straightforward to extend the approach to other pollutants.

Figure 3: Air quality related tweets[6]

After defining all crucial elements of our study **Figure 4** depicts a more abstract view of our general idea. Concretely, bad air quality affects cities and their citizens. Concerned citizens post air quality related content in social media and specifically Twitter. We utilize this information to make predictions about current air quality status in other nearby cities.

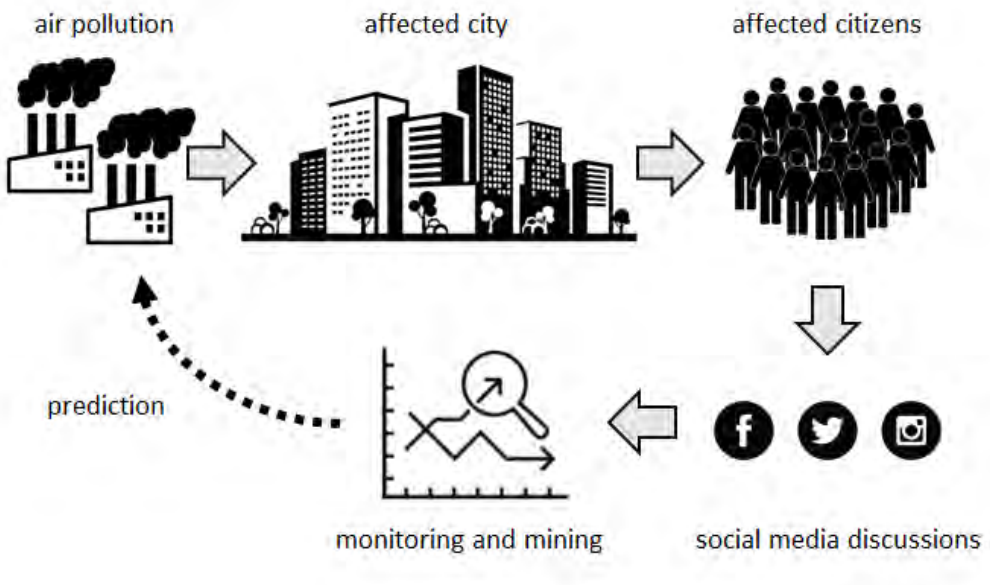

Figure 4: Estimating air quality using Twitter

[6] Source: www.twitter.com

For our work we focused on some major cities that are generally nearby with each other in order to have similar (to some extend) air pollution conditions. We also wanted these cities to have generally high populations aiming to mine as many tweets as possible. In the case of English cities we choose London, Birmingham, Liverpool, Leeds and Manchester. In terms of American cities we focus or research in New York, Philadelphia, Baltimore, Boston and Pittsburgh. Table 1 shows information about the population and the average number of tweets mined per day in each city.

Table 1: City information

Country	**City**	**Population**	**#tweets per day**
UK	London	8.8M	3972
	Liverpool	0.5M	108
	Manchester	0.5M	321
	Birmingham	1.1M	198
	Leeds	0.75M	112
US	New York	8.5M	2564
	Boston	0.7M	574
	Philadelphia	1.6M	478
	Baltimore	0.6M	394
	Pittsburgh	0.3M	169

The developed framework consists of three main components: a) data collection, b) feature extraction, c) multi-task learning. Data collection and feature extraction are described in sections 3.2 and 3.3, respectively, while section 3.1 gives the problem formulation and presents our multi-task learning approach. Section 3.4 briefs the regression algorithm that is used for building the models and performing estimations.

3.1 Problem formulation

Spatial prediction deals with the problem of estimating a quantity of interest in a set of locations, on which the quantity is not measured, based on measurements of the quantity in another set of monitored locations. In this context, the quantity of interest is city-wise average $PM_{2.5}$. C_M/C_U correspond to monitored/unmonitored cities. Due to the high correlation between $PM_{2.5}$ values in nearby locations, the problem can be tackled using simple spatial interpolation techniques such as IDW, which estimates $PM_{2.5}$ in an unmonitored city as a weighted average of the observed $PM_{2.5}$ in nearby cities. In this work we follow a model-based approach. For each city $c_j \in C_M$we construct a set of N training examples $D_{cj} = \{(x^1, y^1), \dots, (x^N, y^N)\}$ where $x^i \in R^D$is a d dimensional input vector that provides an informative summary of the tweets referring to c_j during the i-th temporal bin (see section 0) and $y^i \in R$ is the average $PM_{2.5}$ concentration in c_j during the same temporal bin. For cities $c_q \in C_U$, only x_i are available and our aim is to build a model $h_{c_q} : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ for each $c_q \in C_U$ in order to estimate the unknown y^i. The problem at hand can be considered as a special type of transfer learning [23] where there are multiple target learning tasks $h_{c_q}, c_q \in C_U$ (as in multi-task learning [24]) for which labelled data are completely unavailable (i.e. unmonitored cities) while there are plenty of training data for a number of auxiliary tasks (nearby cities with air quality measurements) $h_{c_j}, c_j \in C_M$ that are related to the target tasks.

3.1.1 Transfer learning approach

Assuming that air pollution exhibits a similar statistical dependence with Twitter activity in cities that share common characteristics (i.e.$P(Y^{c_j}|X^{c_j}) \approx P(Y^{c_i}|X^{c_i})$ as $sim(c_j, c_i) \approx 1$ in case c_j, c_i belong to the same country) we follow a data pooling approach and train a regression model h on $D = \cup_{c_j \in C_M} D_{c_j}$ that learns to simultaneously minimize the prediction error on all monitored cities and we therefore expect it to yield accurate predictions for the unmonitored cities as well.

Besides data pooling, we also apply explicit feature selection to ensure that the learned model will be constrained to a subset of the Twitter-based features that are highly correlated with $PM_{2.5}$ in all cities. To this end, we compute the Pearson correlation[7]

[7] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

coefficient between each feature X_i and the target Y and keep the k features that exhibit the highest correlation. This lower dimensional feature representation is expected to facilitate learning a more robust, city-invariant model.

Finally, under the assumption that the smaller the distance $d(c_j, c_i)$ between two cities, the higher the similarity of their conditional distributions, we develop a weighted data pooling variant where each training example gets a weight that is inversely proportional to the distance between the city it belongs to and the target city. In other words, our model counts more on data from nearest cities to make its predictions rather than the further ones.

3.2 Data collection

There are two types of data needed in order to build our models. The first type of data is Twitter data. Concretely, Twitter data consists of tweets[8] (public posts on Twitter). Tweets are also known as "status updates" of Twitter users. Tweets contain numerous fields, either visible or invisible to users. These fields could be the text of the tweet, user related information, creation time etc. The other type of data is ground truth $PM_{2.5}$ measurements for the selected cities.

3.2.1 Twitter data

Twitter provides a free API that offers real-time access to a sample of its public data. There are two main methods to retrieve tweets using the API. The "location-based" method, allows retrieval of geotagged tweets around an area of interest while the "keyword-based" method retrieves tweets containing specific keywords regardless of location. Some of the previous works that used Sina Weibo as the source OSN, used the location-based method. In Twitter, however, only a tiny fraction of the posts are geotagged (1.5% according to [25]), which significantly limits the number of tweets about air quality that can be collected. In preliminary experiments we found that, e.g., only about 10 air quality related tweets per day are retrieved in London.

[8] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html

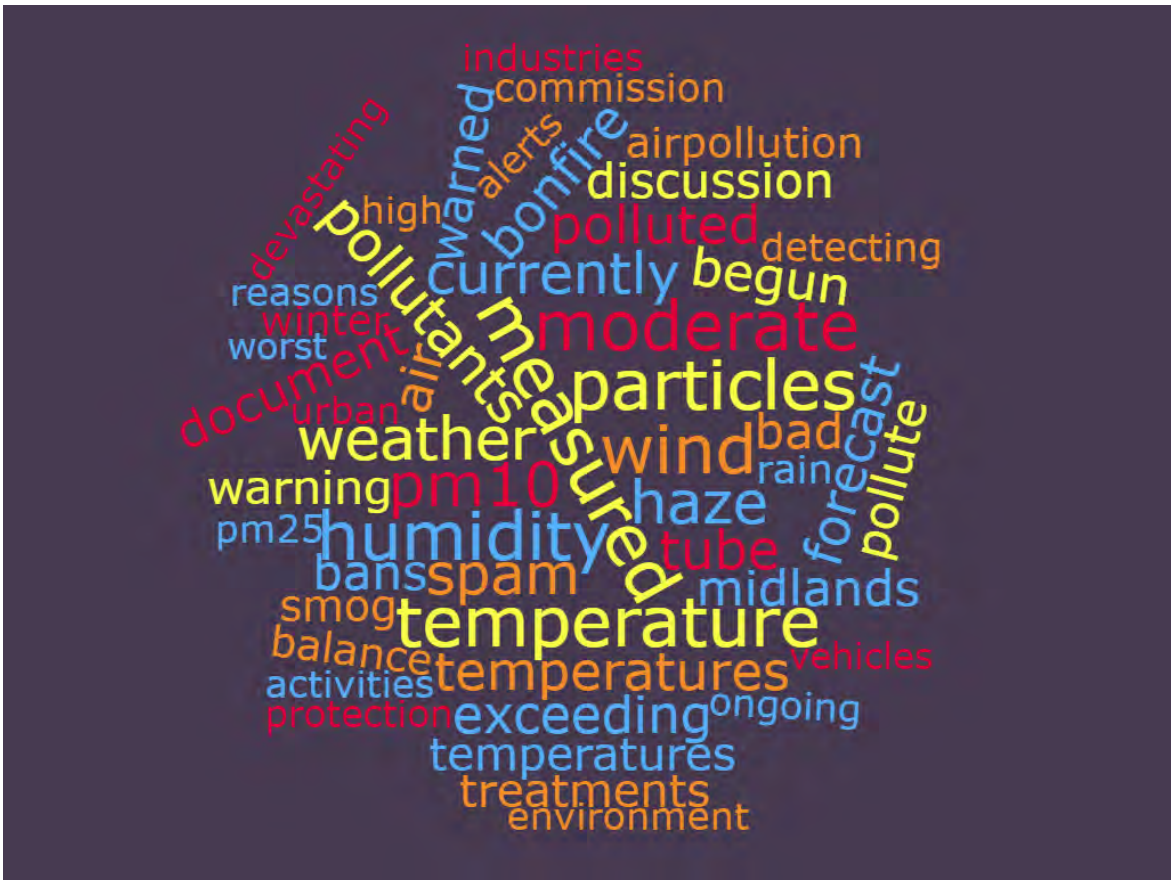

Figure 5: Word cloud with most correlated tweet words with $PM_{2.5}$ values

Therefore, we applied a data collection approach that combines keyword-based search with location inference. Specifically, we track a list of 120 English air quality-related keywords that was composed with the help of air quality experts and store all the returned tweets. Table 8 in **Error! Reference source not found.** contains the list of keywords. Figure 5 shows the most correlated tweet words with $PM_{2.5}$ values in a word cloud. As expected, words that are related with air quality, weather conditions and environment are appeared in the cloud. There are also some irrelevant words, which exhibit a correlation trend with the pollutant probably occurred by chance.

Since the vast majority of the collected tweets are not geotagged, tweet location should somehow be inferred in order to identify tweets related to a city of interest. To this end, previous works simply use the account's declared location as the post's location, assuming that the two locations will coincide in most cases. Here, we follow a more elaborate approach that employs a recent state-of-the-art location estimation method.

According to this approach, the earth surface is divided into (nearly) rectangular cells with sides 0.01° for both latitude and longitude (corresponding to a geodesic length of approximately 1km near the equator), and the term-cell probabilities Figure 6 are computed based on the user count of each term in each cell, based on a training set comprising of the union of the ≈5M training items provided for the 2016 placing task [26] and all geotagged items (≈40M) of the YFCC100M dataset. Given a query text, the most likely cell is derived from the summation of the respective term-cell probabilities. On top of this basic idea, the method features several refinements such as text pre-processing, feature selection, feature weighting, use of multiple resolution grids, etc. More details about these refinements can be found in the original paper [10]

Figure 6: – Illustration of example term-cell probabilities calculated for the grid containing the city of New York

Using this approach, given an item with unknown location, a probability is computed for each cell based on the item's terms and the center of the most likely cell is used as the item's estimated location. Experiments conducted in [27] (section 2.1.2) show that when the confidence of the method for an item's estimated location is higher than 0.6, this location lies within 10 km of the actual location 94% of the time.

In our task, the location estimation method of [10] is applied as follows. Since we are actually interested in the location that the tweet refers to instead of the upload location, we first check if a location can be estimated with high confidence (≥ 0.8) based on the tweet content and in case it does we use it as the tweet location. Otherwise, similarly to previous works, we use the account's declared location as the tweet location. However, instead of relying on simple text matching (which would preclude location

recovery in case of location descriptions referring to, e.g., well-known city districts), we again perform location estimation using the account's location description as input.

To further validate this method we run a simple experiment. After we mine tweets from Twitter API for some period of time. We employ the location estimation method to retrieve only tweets from London. Then we keep only tweets that have geolocation information from twitter which are only a small fraction of all tweets mined. Figure 7 shows that the top 10 real locations of geolocated tweets are indeed mapped to London. In this experiment more than 95% tweets were mapped to London correctly.

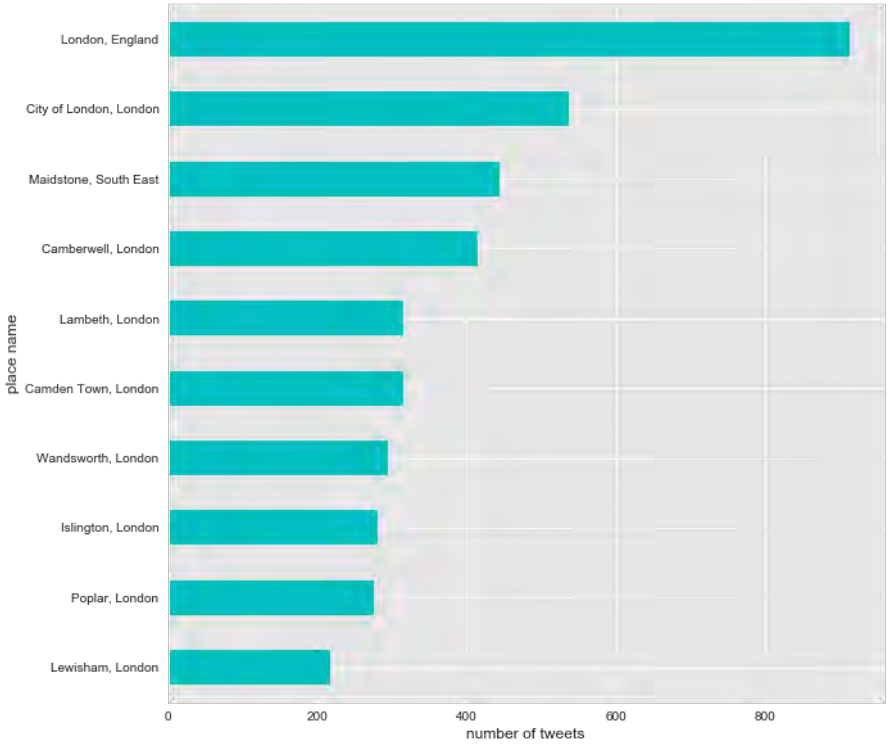

Figure 7: Top 10 real tweet locations assigned to London

The average number of tweets assigned daily to each city using this approach is shown in Table 1. Tweets are stored in a mongoDB[9] database and are mined using twitter4j[10], a Java library for the Twitter API. Figure 8 shows the tweet fields as they are stored in the database.

[9] https://www.mongodb.com/
[10] http://twitter4j.org/en/

(1) 829359820188819457	{ 39 fields }
_id	829359820188819457
createdAt	Feb 8, 2017 6:02:44 PM
id	829359820188819457
text	RT @CleanAirLondon: Next breaches of NO2 hourly legal limit?: #HangerLane @EalingCouncil 15/18 .
source	Twitter for iPhone
isTruncated	false
inReplyToStatusId	-1
inReplyToUserId	-1
isFavorited	false
isRetweeted	false
favoriteCount	0
retweetCount	0
isPossiblySensitive	false
lang	en
contributorsIDs	[0 elements]
retweetedStatus	{ 23 fields }
userMentionEntities	[3 elements]
urlEntities	[1 element]
hashtagEntities	[2 elements]
mediaEntities	[0 elements]
extendedMediaEntities	[0 elements]
symbolEntities	[0 elements]
currentUserRetweetId	-1
user	{ 37 fields }
quotedStatusId	-1
classifiedWP3	Irrelevant
classifiedWP6	Air_pollution
classifiedSH	Irrelevant
wp3	Irrelevant
wp6	Air_pollution
pollution	undefined
confidence	0.97637796
geotagging	tweet_text
loc	{ 2 fields }
location	[2 elements]
near_entity	City of London, United Kingdom
wp3_classifier	Irrelevant
wp6_classifier	Air_pollution
pollution_classifier	undefined

Figure 8: A stored tweet in MongoDB

Figure 9 shows the number of tweets gathered in every 24 hour timestep. We notice that big cities like London and New York have much greater amounts of tweets. Observing the graphs one can also notice some big spikes which may refer to some alarming air quality incidents. These spikes could also be irrelevant with air quality because tweets which are mined with air quality related keywords could possibly be irrelevant with air pollution. For example they keyword “air” could retrieve numerous tweets which are irrelevant with air quality conditions. Additionally, gaps in the plots can indicate two things. The first is that at a current point in time Twitter API was unavailable. As a result no tweets could be retrieved at that time. The second is that the computer which was running the mining process was offline for some reason (internet connection, update restart etc.)

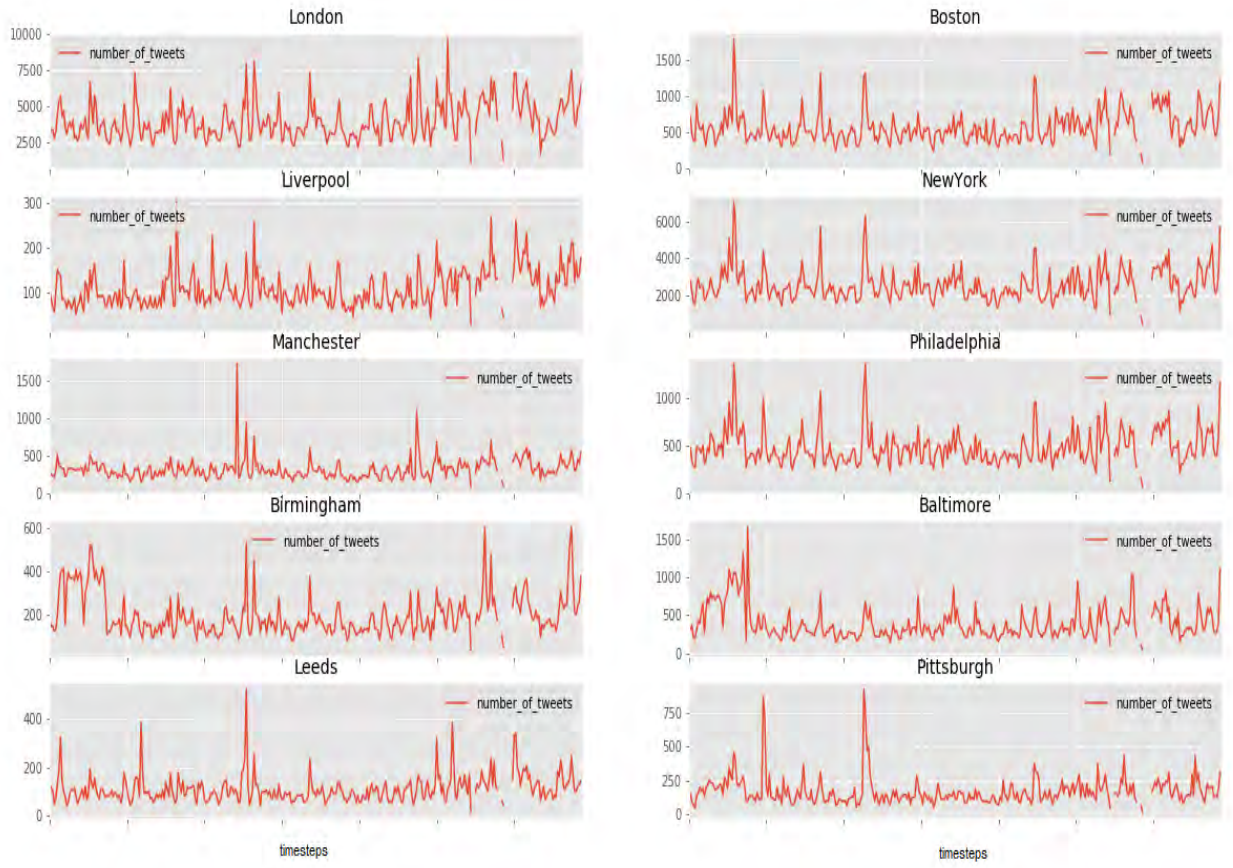

Figure 9: Number of tweets mined for every 24h timestep

3.2.2 Air quality data

To collect ground truth $PM_{2.5}$ measurements for the selected cities we use the OpenAQ[11] platform. OpenAQ is the world's first open, real-time and historical air quality platform, aggregating government-measured and research-grade data - entirely open-source. OpenAQ community is working to fight air inequality - the unequal access to clean air and strives to be a connector for other local organizations and individuals around the world working in their communities to fight air pollution. In addition OpenAQ aims to highlight the role governments and researchers can play in opening up air quality data.

The seed of the idea for OpenAQ emerged from a small open air quality project in Ulaanbaatar, Mongolia, launched by co- founders Joe Flasher and Christa Hasenkopf along with Mongolian colleagues. Amazed at the outsized-impact a little open air quality data can have on a community they started developing OpenAQ. Today, anyone who is using data aggregated from the OpenAQ system, helping build the platform or building on

[11] https://openaq.org

top of the platform in some fashion, and/or interacting with our community around air inequality is part of the OpenAQ Community.

The broad vision of OpenAQ is to empower communities to end air inequality, and the main mission is to enable science, impact policy and empower the public to fight air pollution through open data, open-source tools and cooperation.

So in order to retrieve hourly historical measurements from all stations located within each we attempted to use the OpenAQ API[12]. The main limitation of this approach is that one can retrieve air quality information up until 90 days before the current day. Fortunately, OpenAQ provides a full historical record[13] available for download. This record is contains csv files corresponding to each day of the year for at least 3 years back. Each file contains hourly measurements of a lot of pollutants like $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, CO. These measurements originate from geolocated pollution stations from many cities around the world.

The procedure we follow to collect the $PM_{2.5}$ measurements is described next. Initially, for each city we define its bounding box. To accomplish this we use Flickr Geo API Explorer[14]. Figure 10 shows the bounding box information provided by Flickr.

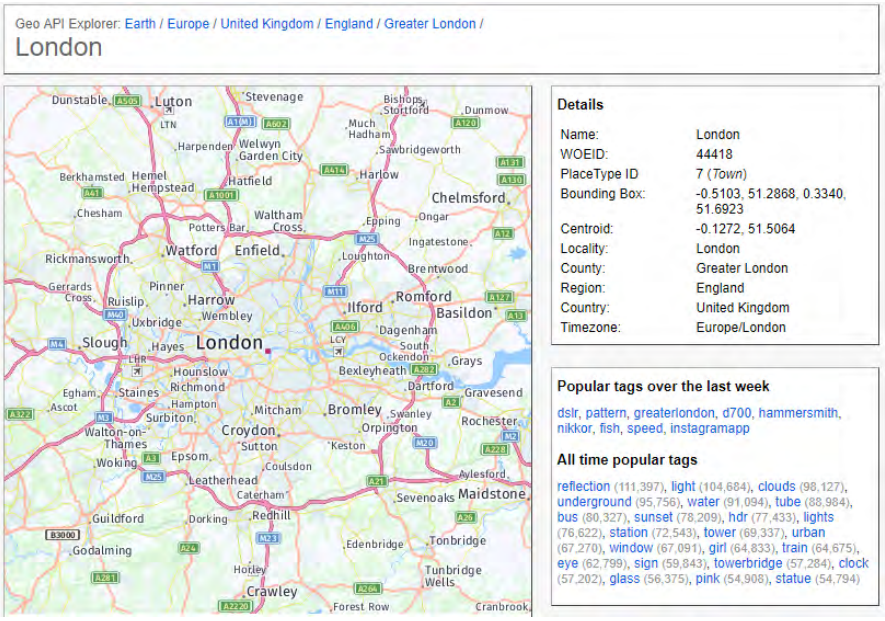

Figure 10: Flick Geo API for the city of London

[12] https://docs.openaq.org/
[13] https://openaq-data.s3.amazonaws.com/index.html
[14] https://www.flickr.com/places

Then we process the csv files with the air pollution measurements and we keep only measurements referring to $PM_{2.5}$ values from stations that are located in the bounding box of the city of interest.

Table 2: Air pollution measurements statistics

City	Number of stations	Average $PM_{2.5}$ (μg/m3)
London	9	11.8
Liverpool	2	7.6
Manchester	3	9.2
Birmingham	2	10.2
Leeds	2	10.1
New York	10	7.9
Boston	4	8.1
Philadelphia	3	10.0
Baltimore	2	8.5
Pittsburgh	2	10.7

The number of stations measuring $PM_{2.5}$ and the average $PM_{2.5}$ values in each city is shown in Table 2. To calculate a single hourly $PM_{2.5}$ value for each city, we average the measurements of the respective stations.

After some careful inspection to the ground truth dataset of $PM_{2.5}$ measurements we noticed that in some cases these measurements return strange numbers and they are probably false measurements. To remedy this we ignore outlier measurements that are 2.5 standard deviations further from the mean of all station measurements in a certain spatiotemporal bin (c, t).

Figure 11 shows the $PM_{2.5}$ values in $\mu g/m^3$ retrieved hourly and aggregated in 24 hour timesteps. As mentioned before aggregation is achieved by averaging the measurements of all city stations and exclude outlier values. Observing the graphs one

can also notice some big spikes which may refer to some alarming air quality incidents. Additionally, gaps in the plots appear because these measurements were not available by OpenAQ API due to station maintenance or some unspecified error.

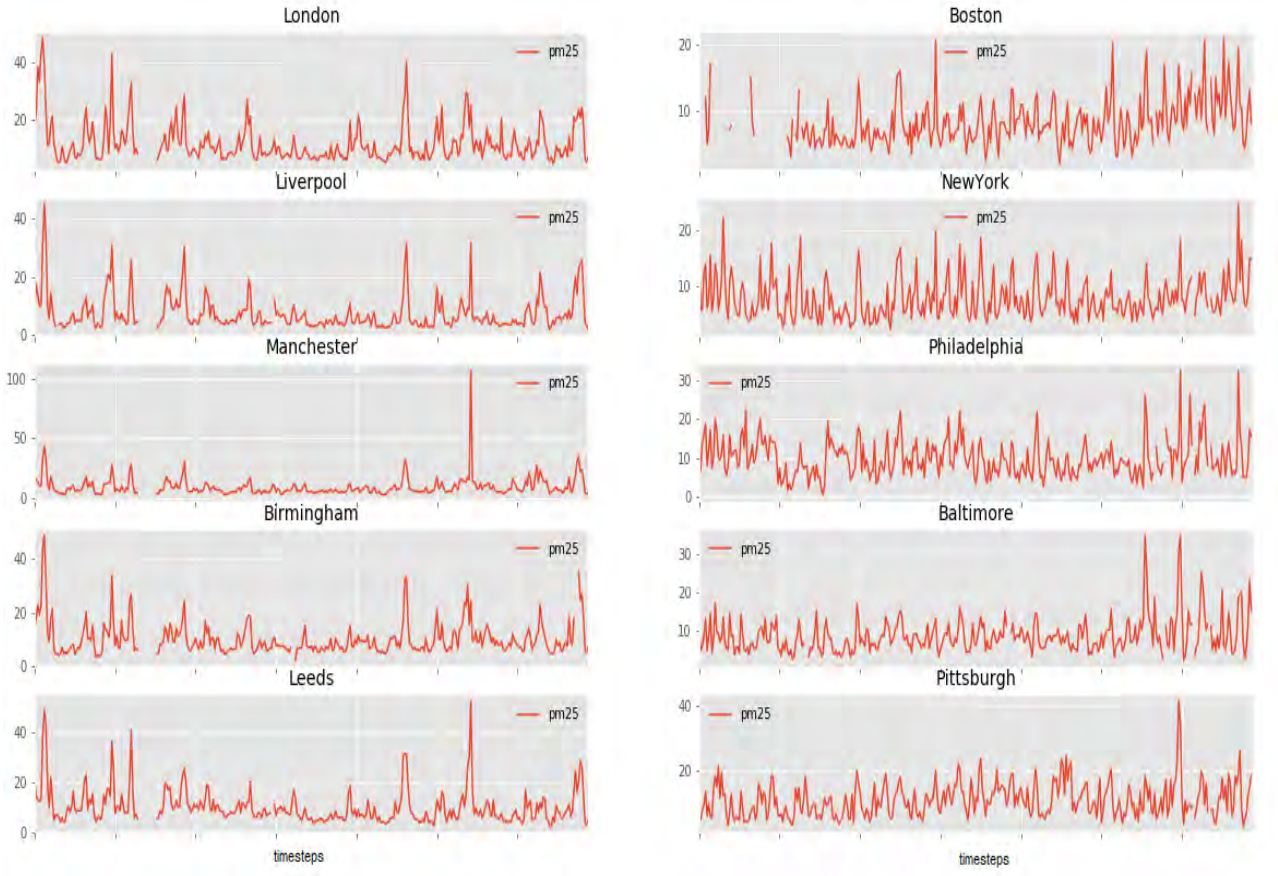

Figure 11: $PM_{2.5}$ measurements from OpenAQ aggregated in 24h timestep

3.3 Feature extraction

To generate a descriptive representation of the tweets assigned to a city c during a temporal bin t (i.e. spatiotemporal bin (c, t)), we use a BoW scheme. First, all tweets are preprocessed by applying tokenization, lowercasing and stopword removal. Then, we create a vocabulary $W = \{w_1, \dots, w_n\}$ that consists of the $n = 10000$ most frequently occurring words in a random 1 million sample of the collected tweets. Using this vocabulary, a BoW vector $x = [x_1, \dots, x_n]$ is generated to represent all tweets in (c, t), where x_i denotes the number of tweets containing w_i divided by the total number of tweets in (c, t). Figure 12 demonstrates a BoW vector example. Each element of the vector corresponds to a single word and the value is the number of tweets that contain this particular word divided by the number of tweets in a temporal bin.

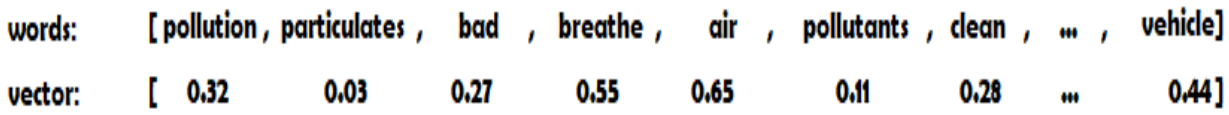

Figure 12: Bag of words representation example

In addition to this "current" BoW representation, we also generate lagged BoW representations (denoted as BoW^{-j}), where instead of considering only the tweets posted during the current temporal bin t we also consider the tweets of the j previous bins. Lagged BoW representations aim at capturing dependencies between Twitter-posts and air pollution that extend beyond the current temporal bin.

Alongside with these feature vectors we also generated some additional features. These features are generated by utilizing the outputs of different tweet classifiers to obtain estimates of the numbers of tweets (in each timestep) that: a) discuss about air quality (Although tweets are collected with air quality related keywords, still many tweets irrelevant with air quality are retrieved) b) provide information about current air quality c) refer to high air pollution. Some details about the classifiers used to generate the above features are provided in

Table 3. The first classifier (AQ_Discussion) is built using a training set of 600 tweets, manually labelled as relevant/irrelevant with air quality. The second classifier (AQ_Sense) is built using a training set of 600 tweets, manually labelled with respect to whether they provide information about current air quality levels (relevant) or not (irrelevant). The third classifier (AP_High) is built using a training set of 200 tweets, manually labelled with respect to whether they refer to high current air pollution levels (relevant) or not (irrelevant). In all cases, tweets were preprocessed by applying stemming and stopword removal and represented using a term frequency bag-of-words representation. As

classification algorithm, we used L2-regularized L2-loss Support Vector Machine [19] (the LibLinear[15] implementation) with default parameters.

Table 3: AQ_General, AQ_Sense and AP_High classifier details

Classifier	#examples	#relevant	#irrelevant	Precision	Recall
AQ_Discussion	600	350	250	90.6%	91.2%
AQ_Sense	600	200	400	88.4%	89.9%
AP_High	200	100	100	80.4%	81.7%

In **Table 3**, we can also see the classification performance of the three classifiers in terms of precision and recall (measured by applying 10-fold cross-validation on the respective training set). We notice that, in all cases, both precision and recall are higher than 80%. This suggests that the estimates generated using the outputs of these classifiers will be quite representative of the actual numbers.

Table **9** and Table 10 of the **Error! Reference source not found.**, show examples of tweets classified as relevant by the AQ_Sense and the AQ_Discussion classifier respectively. Table 11 shows examples of tweets classified as relevant by the AP_High classifier.

[15] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

In Figure 13 we plot the daily numbers of London tweets for a random period of 3 months, classified as relevant/irrelevant by the AQ_Discussion classifier (AQ_D/Irrelevant), classified as relevant by the AQ_Sense classifier (AQ_S) and classified as relevant/irrelevant by the AP_High classifier (High/Irrelevant). We notice that a considerable number of tweets irrelevant with air quality are retrieved. Moreover, we see that, as expected, the number of AQ_S tweets is always smaller than the number of AQ_D tweets and that the majority of AQ_S tweets are classified as related to high air pollution. This can be explained by the fact that the list of keywords used for querying the Twitter API is primarily oriented towards bad air quality conditions and by the fact that people tend to tweet more about bad air quality.

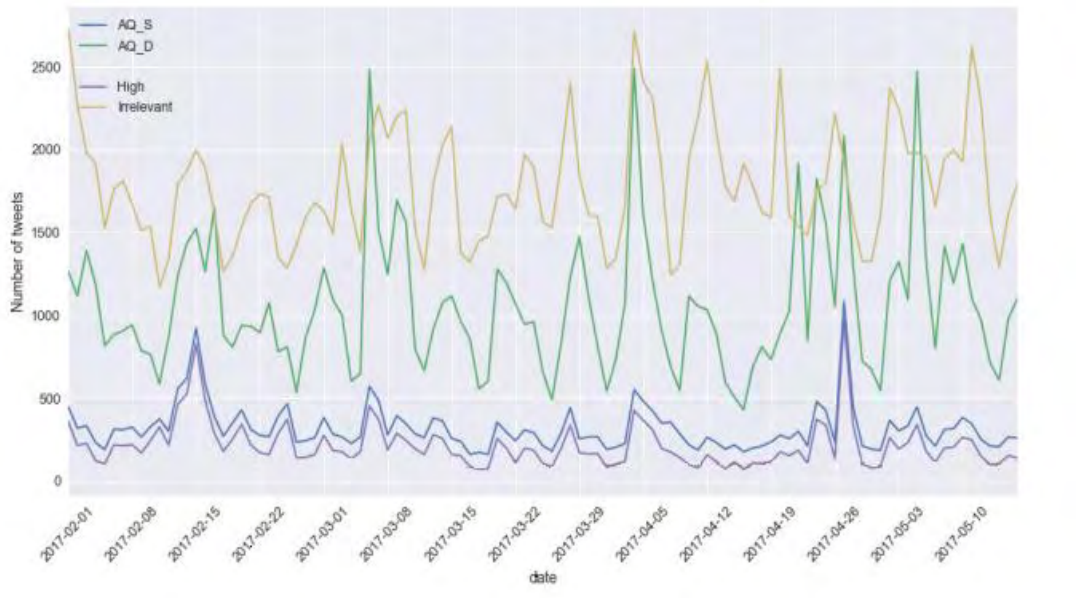

Figure 13: Daily numbers of AQ_D, AQ_S, High and Irrelevant tweets in London. Irrelevant refers to tweets classified as irrelevant by the AQ_Discussion classifier.

3.4 Regression models

After collecting the data needed to create our models we need to choose a regression algorithm to train them. We tried various types of regressors like Support Vector regression (SVR) [19], Lasso regression [28], Ridge regression [18], Random Forest regression [29] and Gradient Boosting regression [21]. Gradient Boosting regression was found to perform equally good or better compared to other algorithms in a set of

preliminary experiments[16]. As a result we use this regression algorithm to build our models.

3.4.1 Gradient Tree Boosting

Gradient Tree Boosting or Gradient Boosted Regression Trees (GBRT) is a generalization of boosting to arbitrary differentiable loss functions. GBRT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems. Gradient Tree Boosting models are used in a variety of areas including Web search ranking.

- The advantages of GBRT are:
- Natural handling of data of mixed type (= heterogeneous features)
- Predictive power

Robustness to outliers in output space (via robust loss functions)

The disadvantages of GBRT are:

- Scalability, due to the sequential nature of boosting it can hardly be parallelized.

GBRT considers additive models of the following form:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Where $h_m(x)$ are the basis functions which are usually called weak learners in the context of boosting. Gradient Tree Boosting uses decision trees of fixed size as weak learners. Decision trees have a number of abilities that make them valuable for boosting, namely the ability to handle data of mixed type and the ability to model complex functions.

Similar to other boosting algorithms GBRT builds the additive model in a forward stagewise fashion:

$$F_m = F_{m-1}(x) + \gamma_m h_m(x)$$

At each stage the decision tree $h_m(x)$ is chosen to minimize the loss function **L** given the current model F_{m-1} and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + arg\min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x))$$

[16] https://github.com/MKLab-ITI/twitter-aq

The initial model $F(0)$ is problem specific, for least-squares regression one usually chooses the mean of the target values.

Gradient Boosting attempts to solve this minimization problem numerically via steepest descent: The steepest descent direction is the negative gradient of the loss function evaluated at the current model F_{m-1} which can be calculated for any differentiable loss function:

$$F_m(x) = F_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i))$$

Where the step length γ_m is chosen using line search:

$$\gamma_m = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}$$

The algorithms for regression and classification only differ in the concrete loss function used. For our case the least squares lost function was used.

3.4.2 Least squares

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least-squares problems fall into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The nonlinear problem is usually solved by iterative refinement; at

each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Polynomial least squares describe the variance in a prediction of the dependent variable as a function of the independent variable and the deviations from the fitted curve.

The objective consists of adjusting the parameters of a model function to best fit a data set. A simple data set consists of n points (data pairs) $(x_i, y_i), i = 1, \ldots, n$ where x_i is an independent variable and y_i is a dependent variable whose value is found by observation. The model function has the form $f(x, \beta)$, where m adjustable parameters are held in the vector β. The goal is to find the parameter values for the model that "best" fits the data. The fit of a model to a data point is measured by its residual, defined as the difference between the actual value of the dependent variable and the value predicted by the model:

$$r_i = y_i - f(x_i, \beta)$$

The least-squares method finds the optimal parameter values by minimizing the sum, S, of squared residuals:

$$S = \sum_{i=1}^{n} r_i^2$$

CHAPTER 4

EXPERIMENTS

4.1 Experimental setup

To simulate the spatial air quality prediction task, we collected data for five cities in the UK and five cities in the US for a time period spanning almost a whole year (8/2/2017-18/1/2018). Each city is in turn treated as the test city (hypothetically without air quality measurements) and all the remaining neighbouring cities are used for training. For each city, we train and evaluate models able to perform predictions at three different temporal granularities: 6, 12 and 24 hours. This is accomplished by grouping the hourly $PM_{2.5}$ observations into correspondingly sized temporal bins and calculating a single ground truth $PM_{2.5}$ value for each bin as the average of the hourly values. In some cases measurements from ground truth stations are missing. We deal with this by ignoring the all the missing values and we average the valid measurements. If there are no valid measurements in a temporal step then we remove this from the dataset. Prediction accuracy for each city and temporal granularity is measured in terms of Root Mean Squared Error (RMSE) and macro averaging is applied to calculate country-wise or overall performance (denoted as aRMSE). In all our experiments we use Gradient Tree Boosting as the regression algorithm, since it is recognized as one of the best off-the-shelf supervised learning algorithms [30] and was found to perform equally good or better compared to other algorithms in a set of preliminary experiments.

Experiments carried out using python programming language and sklearn[17] machine learning library. The complete code can be found in the following link: https://github.com/MKLab-ITI/twitter-aq. Apart from the code this link contains the datasets in a sparse vectorised format for every city and granularity.

[17] http://scikit-learn.org

4.2 Baseline performance

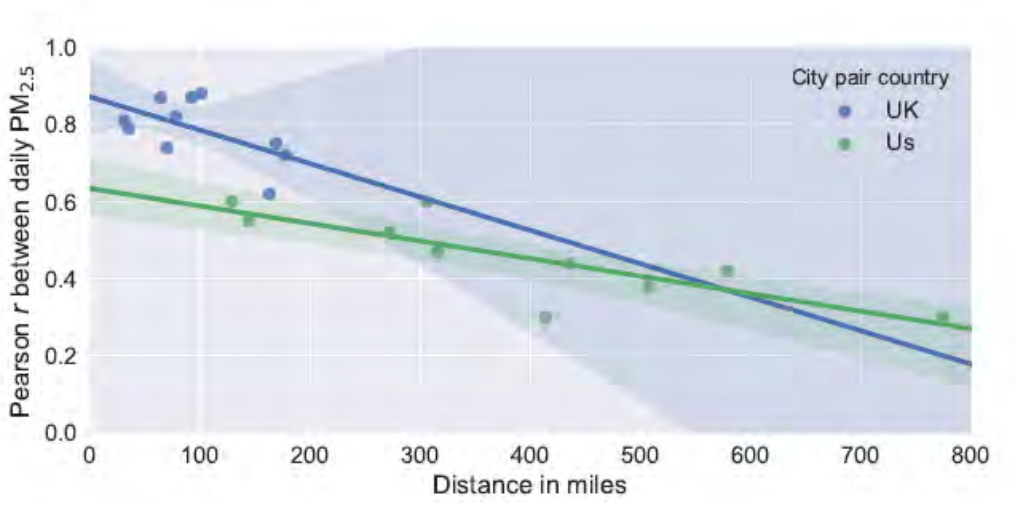

Figure 14: Scatter plot of distances and Pearson r between average daily $PM_{2.5}$

Figure 14 shows a scatter plot of the distances and the Pearson r between average daily $PM_{2.5}$ concentrations for all distinct city pairs in UK and US. Clearly, the smaller the distance between two cities, the higher the correlation between their average daily $PM_{2.5}$ concentrations. Given this high spatial dependence, it is not surprising that spatial interpolation methods such as IDW yield highly accurate estimates as shown in Table 4. Moreover, we see that a baseline that always predicts the mean $PM_{2.5}$ value per city, although it has worse performance than IDW, it has a relatively small error which is due to the fact that $PM_{2.5}$ levels in the studied cities are generally low and exhibit small variability. Another thing that we can observe from Table 4 is that in larger temporal granularities the error tends to become smaller. This behaviour can be explained because in larger granularities there are more averaged data and make the prediction task less strict.

Table 4: aRMSE of baseline methods

	UK			US			Overall		
	6h	12h	24h	6h	12h	24h	6h	12h	24h
IDW	3.79	3.34	3.09	4.12	3.73	3.73	3.96	3.54	3.25
Mean	7.00	6.11	6.36	4.60	4.26	4.02	5.80	5.46	5.19

4.3 Within-city predictions

Before evaluating spatial $PM_{2.5}$ prediction using our transfer learning approach, we first evaluate the predictability of $PM_{2.5}$ in each city using a model trained on Twitter and ground truth data of the city. As already discussed, this represents an unrealistic setup because ground truth data is not available for unmonitored cities. However, it is suitable for assessing the effectiveness of different Twitter features. In this set of experiments, data from each city is split based on time, using odd months for training and even months for testing.

Table 5: aRMSE of different Twitter features

	#tw	**#aqs**	**#high**	**All**	**BoW**	**BoW^{-1}**	**BoW^{-2}**
6h	5.96	5.93	5.98	5.84	5.15	4.99	**4.97**
12h	6.17	5.98	6.02	5.77	4.96	**4.84**	5.16
24h	5.83	6.11	5.82	5.52	**4.65**	4.96	5.16

Table 5 shows the results obtained using models trained on "current" and lagged BoW features, as well as four simpler Twitter feature. Some of them were presented in section 0 : '#tw' (total number of tweets in each spatiotemporal bin), '#aqs' (number of tweets that provide information on current air quality), '#high' (number of tweets that refer to high air pollution levels) and 'all' (the concatenation of '#tw', '#aqs' and '#high'). Note that in these experiments we exclude #aqd feature (number of tweets that refer to general air quality discussions) as it scored poorly in preliminary experiments. We notice that for all temporal granularities, 'all' leads to better accuracy than '#tw', '#aqs' and '#high', suggesting that these features capture complementary information about current air quality. However, we see that the best performance for each temporal granularity is obtained by a BoW variant and, interestingly, we notice that for finer temporal granularities it is beneficial to use lagged BoW features (BoW^{-2} and BoW^{-1} for the 6- and the 12-hour temporal granularity, respectively). Based on these results, subsequent

experiments employ the best performing BoW representation for each temporal granularity.

4.4 Cross-city predictions

We now turn into the main focus of our research, i.e. spatial $PM_{2.5}$ prediction, and evaluate our transfer learning approach according to the setup described in section 0. Table 6 shows the results obtained when using full-dimensional BoW vectors ('full' column) as well as vectors where only the top-k most correlated features are kept, with (w=1) and without (w=0) sample weighting. First, we observe that the performance of full-dimensional BoW is considerably worse compared to the within-city setup. As expected, the absence of city-specific training data makes the learning task more difficult. With respect to the different transfer learning setups, we see that joint feature selection results in important performance gains in all temporal granularities, with the best results obtained when the top 50 or 100 features are used. As we mentioned in 0 the top features are selected by calculating their Pearson correlation with ground truth $PM_{2.5}$ measurements. This will ensure that the most descriptive features will be used to train the regression models.

Sample weighting, on the other hand, has a less pronounced but consistently positive effect.

Table 6: Cross-city aRMSE with different transfer learning setups

		full	**k=10**	**k=20**	**k=50**	**k=100**	**k=200**	**k=500**
w=0	6h	5.36	5.48	5.28	**5.21**	5.24	5.29	5.31
	12h	5.21	5.29	5.18	5.12	**5.09**	5.11	5.15
	24h	4.97	4.89	4.78	4.78	**4.75**	4.79	4.86
w=1	6h	5.35	5.47	5.27	**5.21**	5.24	5.29	5.16
	12h	5.21	5.26	5.18	5.11	**5.08**	5.11	5.16
	24h	4.95	4.85	4.77	4.76	**4.73**	4.77	4.84

Comparing the performance of our Twitter-based estimates with those of IDW, we notice that they do not perform on par. We believe that this result should be largely

attributed to the fact that the studied cities exhibited very good air quality conditions for an overwhelming part of the studied period which makes it less likely for people to express their feelings about air quality on Twitter. Our findings match those reported in [17] where IDW was also found more accurate than the proposed approach under good air quality conditions.

As in **Error! Reference source not found.** we notice the same trend that that in larger temporal granularities the error tends to become smaller. In Twitter case this can be easily explained if we take into consideration that 6 hour temporal bins may correspond to late night hours where twitter usage is restricted. This means that for these timesteps predictions not going to be accurate. In 24 hour bins daily twitter information is aggregated leading to more normalized and accurate results.

Figure 15 depicts the prediction of $PM_{2.5}$ values with sample weighting in the city of London. Actual corresponds to $PM_{2.5}$ ground truth measurements. IDW is the baseline prediction and Twitter is the combinations of Twitter features with spatial interpolation. The prediction period on this figure covers approximately two and a half months. We observe that even though IDW offers a fairly good prediction, Twitter improves the performances in some cases.

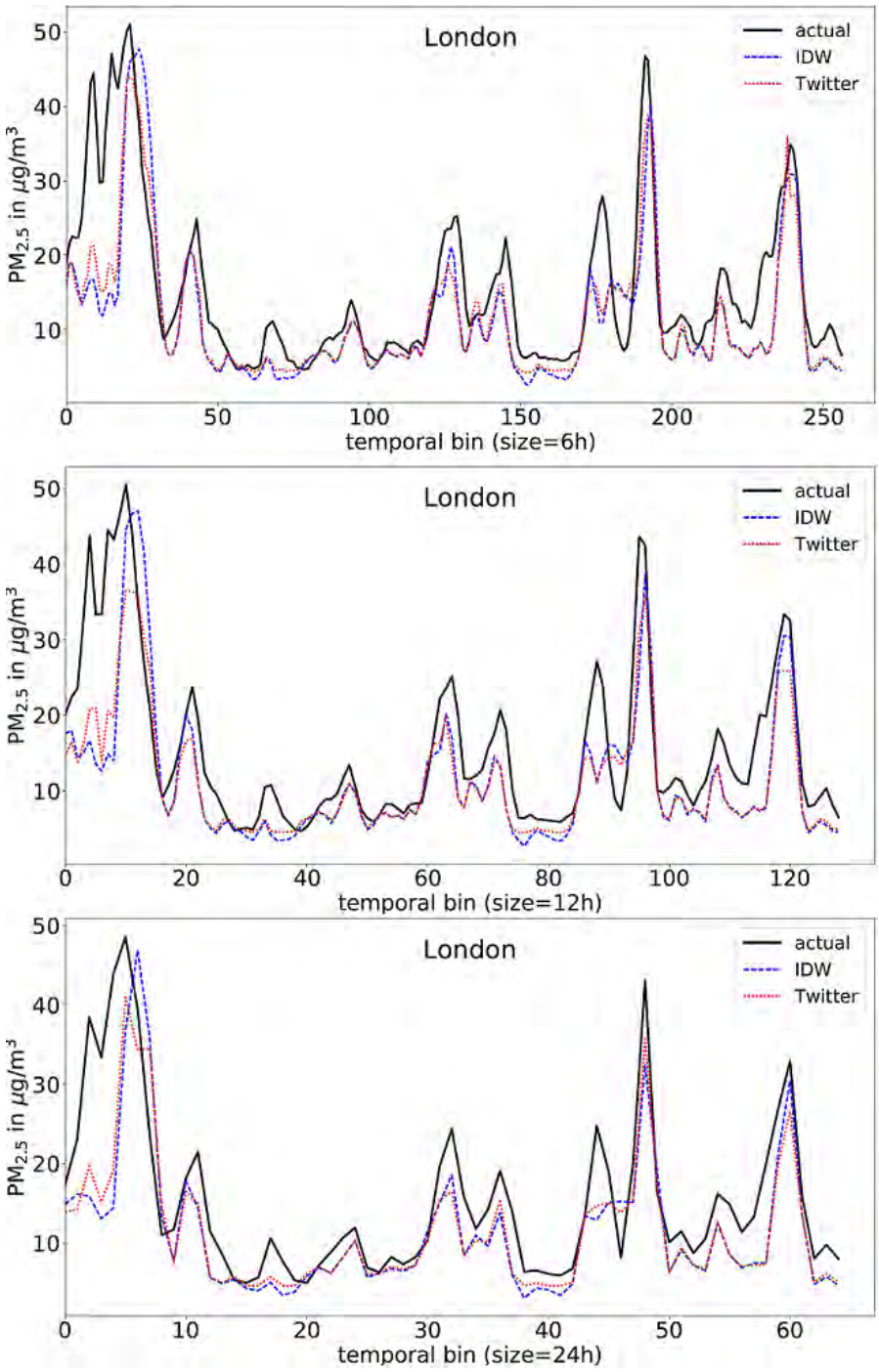

Figure 15: Prediction results on London for 6h, 12h and 24h timesteps

Despite the fact that IDW yields better results, we notice that Twitter-based estimates carry considerable predictive power as they manage to obtain significantly lower error than the mean baseline. Motivated by that, we evaluated a late fusion scheme that combines our Twitter-based estimates with the IDW estimates by learning a meta-model that uses two features: a) IDW estimates for the training cities, b) Twitter-based estimates for the training cities (obtained through inner cross-validation). This model

obtains an aRMSE of 4.15, 4.00 and 3.63 for the temporal granularities of 6, 12 and 24 hours respectively. Although its performance is still worse on average compared to IDW, Table 7 shows that it performs better than IDW in 3 out of 10 cities: Boston, London and Pittsburgh.

Table 7: Per city results

City	IDW error	Twitter with IDW
Baltimore	3.12	3.85
Birmingham	2.28	2.95
Boston	3.34	**3.02**
Leeds	3.32	3.36
Liverpool	3.65	4.01
London	4.66	**4.25**
Manchester	3.12	4.01
New York	3.14	3.37
Philadelphia	3.98	4.2
Pittsburgh	5.16	**4.88**

We notice that these cities are the most distant (on average) to the rest of the studied cities in each country, thus limiting the accuracy of spatial interpolation. This shows that exploiting Twitter information can be beneficial for improving air quality estimates even in cities with good average air quality conditions when they lie far from monitored cities.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this MSc thesis we presented a methodology for performing Twitter-based air quality estimations on cities that lack monitoring infrastructure. Our approach was found to provide fairly accurate estimates on a case study involving cities in the UK and the US. Although these estimates are less accurate than estimates obtained through spatial interpolation, we found that by combining the two types of estimates it is possible to improve accuracy in certain cities.

In the future, we would like to extend our empirical study to additional air pollutants and to a larger and more diverse (in terms of population, country, air quality levels) set of cities. Moreover, we would like to experiment with more sophisticated textual representations (e.g. [31]) and transfer learning methods. Also, we aim to deal with restricted Twitter activity in late night hours which is a factor that harms the overall performance.

Finally, it would be interesting to study whether better accuracy could be obtained by exploiting the image content of tweets using image-based air quality estimation approaches (e.g. [32]).

BIBLIOGRAPHY

[1] EEA, "Air quality in europe 2017," [Online]. Available: https://www.eea.europa.eu/publications/air-quality-in-europe-2017. [Accessed 06 02 2018].

[2] Environmental Protection UK, "Healthy air where you live," [Online]. Available: https://www.healthyair.org.uk/documents/2013/02/healthy-air-community-campaign-pack-2012.pdf. [Accessed 06 02 2018].

[3] Unicef, "Understanding and addressing the impact of air pollution on children's health in Mongolia," [Online]. Available: https://www.unicef.org/environment/files/Understanding_and_addressing_the_impact_of_air_pollution.pdf. [Accessed 06 02 2018].

[4] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday and A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?," *Environment International,* pp. 293-302, 2017.

[5] H. T. Shahraiyni and S. Sodoudi, "Statistical modeling approaches for pm10 prediction in urban areas; a review of 21st-century studies," *Atmosphere,* vol. 7, no. 2, p. 15, 2016.

[6] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris and A. Jaimes, "Sensing trending topics on twitter," *IEEE Transactions on Multimedia,* vol. 15, no. 6, pp. 1268-1282, 2013.

[7] V. Lampos, T. D. Bie and a. N. Cristianini, "Flu detector-tracking epidemics on twitter," *Joint European conference on machine learning and knowledge discovery in databases.,* pp. 599-602, 2010.

[8] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th national conference on World wide web. ACM*, 2010.

[9] J.-M. Xu, A. Bhargava, R. Nowak and X. Zhu, "Socioscope: Spatiotemporal signal recovery from social media," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012.

[10] G. Kordopatis-Zilos and a. I. K. S. Papadopoulos, "Geotagging text content with language models and feature mining," *Proceedings of,* vol. 105, no. 10, pp. 1971-1986, 2017.

[11] W. Jiang, Y. Wang, M.-H. Tsou and X. Fu, "Using social media to detect outdoor air pollution and monitor air quality index (aqi): a geo-targeted spatiotemporal analysis framework with sina weibo (chinese twitter)," *PloS one,* vol. 10, no. 10, p. e0141185, 2015.

[12] S. Wang, M. J. Paul and M. Dredze, "Social media as a sensor of air quality and public response in china," *Journal of medical Internet research,* vol. 17, no. 3, 2015.

[13] Z. Tao, A. Kokas, R. Zhang, D. S. Cohan and D. Wallach, "Inferring atmospheric particulate matter concentrations from chinese social media data," *PloS one,* vol. 11, no. 9, p. e0161389, 2016.

[14] M. L. Stein, "Interpolation of spatial data: some theory for kriging.," in *Springer*

Science & Business Media, 2012.

[15] D. G. Fox, "Judging air quality model performance," *Bulletin of the American Meteorological Society,* vol. 62, no. 5, pp. 599-609, 1981.

[16] G. Hoek, R. Beelen, K. D. Hoogh, D. Vienneau, J. Gulliver, P. Fischer and D. Briggs, "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric environment,* vol. 42, no. 33, pp. 7561-7578, 2008.

[17] S. Mei, H. Li, J. Fan, X. Zhu and C. R. Dyer, "Inferring air pollution by sniffing social media," *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE,* p. 534–539, 2014.

[18] D. W. Marquardt and R. D. Snee, "Ridge Regression in Practice," *American statisticial,* vol. 29, no. 1, pp. 3-20, 1975.

[19] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning,* vol. 3, no. 20, p. 273–297, 1995.

[20] J. D. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[21] J. H. Friedman, "Γreedy function approximation: a gradient boosting," *Annals of statistics,* pp. 1189-1232, 2001.

[22] US Environmental Protection Agency, "What is PM2.5 and Why You Should Care," [Online]. Available: https://blissair.com/what-is-pm-2-5.htm. [Accessed 6 2 2018].

[23] S. J. Pan and Q. Yang, "A survey on tranfer learning," *IEEE Transactions on knowledge and data engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[24] R. Caruana, "Multitask learning," *Machine learning,* vol. 28, no. 1, pp. 41-75, 1997.

[25] F. Morstatter, J. Pfeffer, H. Liu and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," *ICWSM,* 2013.

[26] J. Choi, " The placing task:A large-scale geo-estimation challenge for social-media videos and images," *In Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia,* pp. 27-31, 2014.

[27] E. Spyromitros-Xioufis, S. Papadopoulos, A. Moumtzidou, S. Vrochidis and Y. Kompatsiaris, "hackair deliverable d3.2: 2nd environmental node discovery indexing and data acquisition," July 2017. [Online]. Available: https://www.researchgate.net/publication/324594192 hackAIR_deliverable_D32_2nd_environmental _node_discovery_indexing_and_data_acquisition.

[28] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B ,* vol. 58, no. 1, pp. 267-288, 1996.

[29] L. Breiman, "Random Forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[30] T. H. a. R. T. J. Friedman, "The elements of statistical learning," *Springer series in statistics New York,* vol. 1, 2001.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems,* pp. 3111-3119, 2013.

[32] E. Spyromitros-Xioufis, A. Moumtzidou, S. Papadopoulos, .. K. S. Vrochidis, A. K. Georgoulias, G. Alexandri and K. Kourtidis, "Towards improved air quality monitoring using publicly available sky images," *Multimedia Tools and Applications for Environmental and Biodiversity Informatics,* 2018.

ΠΑΡΑΡΤΗΜΑ

DATA FROM TWITTER

Table 8: Air quality related keywords

Air quality related keywords in English
air pollution, airpollution, airquality, air status, polluted air, noxious gases, fumes, air quality index, air quality, air pollution index, aqi, polluted atmosphere, pollution, pollutants, pollutant, emissions, emission, smog, antipollution, anti-pollution, clean atmosphere, ozone pollution, haze, sky pollution, clean air, clean sky, unpolluted air, beautiful sky, fresh air, air toxins, climate pollution, pollution mask, particulates, particulate matter, pesticides, air particles, wood burning, fuel pollution, coal pollution, sulfur dioxide, carbon monoxide, carbon dioxide, ground level ozone, nitrogen dioxide, co2, co emissions, o3, pm10, pm2.5, no2, so2, ozone depleters, coarse particles, fine particle, fine particles, fine dust, yellow dust, chlorofluorocarbon, chlorofluorocarbons, nitrogen oxide, nitrogen oxides, nitrates, radioactive, peroxyacetyl nitrate, aqi observatory, health hazard, toxic air, filthy air, unhealthy air, toxic smog, toxic emissions, toxic atmosphere, poisonous air, poisonous chemicals, poisonous gases, filthy atmosphere, bad air, dusty air, hazardous, hazards, contamination, asthma, bronchitis, emphysema, pulmonary disease, respiratory, respiratory illness, respiratory problems, respiratory disease, breathing problems, allergic reaction, allergic reactions, cardiac disease, heart disease, cardiovascular, cough, coughing, wheeze, wheezing, health problems, breathe, inhale, lung, lungs, itchy eyes, breathing difficulties, breathing difficulty, pneumonia, hospital admissions, sore throat, phlegm, sputum, breathless, mucus, out of breath, smell pollution, sense pollution, pollution level, severe pollution, environmental pollution

Table 9: Examples of AQ_S tweets in London

AQ_S tweets
RT @Jackie_News: Wandsworth, Kingston and Hounslow are some of the areas under a toxic “red alert” today http://t.co/o1pkytqhJM
RT @ClientEarth: #Knightbridge has become the third monitor breach legal #airpollution limits. Support our fight for #clearnair…
RT @LondonAir: High air pollution forecast valid from Thursday 19 January to end of Thursday 19 January https://t.co/W3D5gD9fxO #airpollution
RT @MayorofLondon: London’s dirty air is a public health crisis. I’m committed to tackling this. Read more about my plans here: https://t.c...
@tony_olmstead @SenSanders no a hypocrite if no alternative. Do people campaigning against air pollution have to stop breathing?
RT @ LondonAir: This pic shows PM2.5 particulate #airpollution building through this cold snap since Sat. Widespread Moderate, like…

Table 10: Examples of AQ_D tweets in London

AQ_D tweets
RT @RSLenvironment: AirPollution is a crisis that “plagues” the UK, particularly children, according to UN human rights expert http://t.c...
Greenhouse Morning News is out! Top stories: EU ‘ready to fight’ for # climate & UK #airpollution crises…
RT @GasNaturally: Another proof that #NatGas reduces #PowerGen CO2 #emissions when used instead of #Coal
Our #future. EVs will help “goal to reduce greenhouse gas #emissions by 80-95% by 2050” https://t.co/L46Xm7VI29
This ‘smog-eating’ city sculpture can combat London’s toxic pollution as effectively as 275 trees https://t.co/99FY7EyiSW

RT @CleanSpaceLDN: Bad night's sleep? #Airpollution could be to blame, study finds https://t.co/MjYQiaVDB7 via @guardian

Table 11: Examples of AQ_High tweets in London

AP_Severity High
RT @PlumeInLondon: High pollution (50) at 10PM. High for #London. Avoid physical activities if sensitive https://t.co/3LVRgps965
London's air pollution is killing me. Coughs now sound like squeaky chew toy. #sendhelp #sendventolin
RT @cargill_taxi: And the mayor of London tries to blame poor air quality on toxic air from German factories. You need to look a bit…
@claireL23 The traffic, poor air quality, the light pollution, the lack of green space, the concrete jungle, the bu… https://t.co/JkXYWA16GM
RT @SkyNews: THE GUARDIAN FRONT PAGE: "Toxic air risk to one in four London schools" #skypapers https://t.co/2c6ANlujep
RT @MayorofLondon: London’s toxic air is a public health emergency. Here is what I’m doing about it https://t.co/YHw2CVepPI